CLAIMS

What is claimed is:

1. A method for managing admission of requests to a streaming media server, the method comprising:

receiving a new request for a streaming media file to be served by a streaming media server;

performing a resource availability check for the streaming media server to determine whether the streaming media server has sufficient available resources to service the new request; and

performing a quality of service guarantee check for the streaming media server to determine whether acceptance of the new request will violate, at any point in the future, a desired quality of service provided by the streaming media server for any previously accepted requests.

2. The method of claim 1 wherein said resource availability check comprises:

using a segment-based memory model to determine whether at least a portion of the requested streaming media file is in the streaming media server's memory.

3. The method of claim 2 further comprising:

determining from the segment-based memory model a cost associated with serving the requested streaming media file from the streaming media server.

4. The method of claim 1 wherein said resource availability check comprises:

determining a cost associated with serving the requested streaming media file from the streaming media server.

5. The method of claim 4 wherein the cost comprises:

a cost of serving the requested streaming media file either from memory or from disk.

6. The method of claim 1 wherein said resources comprise memory resources and disk resources.

7. The method of claim 1 wherein said sufficient available resources to service the new request comprises sufficient resources available so as not to overload the streaming media server.

8. The method of claim 1 wherein said desired quality of service comprises real-time delivery of streaming media files requested by said previously accepted requests.

9. The method of claim 1 further comprising:

if determined that the streaming media server has sufficient available resources to service the new request and determined that acceptance of the new request will not violate, at any point in the future, said desired quality of service provided by the streaming media server for any previously accepted requests, then the streaming media server serving the requested streaming media file for said new request.

10. The method of claim 1 further comprising:

if determined that the streaming media server doe not have sufficient available resources to service the new request or determined that acceptance of the new request will violate, at any point in the future, said desired quality of service provided by the streaming media server for any previously accepted requests, then rejecting the new request for service by the streaming media server.

11. A method for managing admission of requests to a media server, the method comprising:

receiving a new request for a streaming file to be served by a media server;

determining a cost to the media server for serving the requested streaming file, wherein the cost corresponds to the media server's resources to be consumed in serving the requested streaming file; and

determining, based at least in part on the cost, whether to admit the new request for service by the media server.

12. The method of claim 11 wherein said determining said cost comprises:

determining a segment-based memory model that identifies content of the media server's memory as of a time that the new request is received; and

using the segment-based memory model to determine whether at least a portion of the requested streaming file is in the media server's memory.

13. The method of claim 12 wherein the cost comprises:

a cost of serving the requested streaming file from memory if determined that the requested streaming file is in the media server's memory and a cost of serving the requested streaming file from disk if determined that the requested streaming file is not in the media server's memory.

14. The method of claim 11 wherein said determining whether to admit the new request for service by the media server comprises:

performing a resource availability check for the media server to determine whether the media server has sufficient available resources to service the new request.

15. The method of claim 14 wherein said sufficient available resources to service the new request comprises sufficient resources available so as not to overload the media server.

16. The method of claim 14 wherein said determining whether to admit the new request for service by the media server further comprises:

performing quality of service guarantee check for the media server to determine whether acceptance of the new request will violate, at any point in the future, a desired quality of service provided by the media server for any previously accepted requests.

17. A system comprising:

server having a memory, wherein said server is operable to serve at least one streaming file to clients communicatively coupled thereto; and

an admission controller operable to receive a new request for a streaming file to be served by said server, determine a cost to the server for serving the requested streaming file, wherein the cost corresponds to the server's resources to be consumed in serving the requested streaming file, and determine, based at least in part on the cost, whether to admit the new request for service by the server.

18. The system of claim 17 wherein said admission controller is further operable to determine a segment-based memory model that identifies content of the server's memory as of a time that the new request is received, and said admission controller is operable to use the segment-based memory model to determine whether at least a portion of the requested streaming file is in the server's memory.

19. The system of claim 17 wherein the cost comprises:

a cost of serving the requested streaming file from memory if determined that the requested streaming file is in the server's memory and a cost of serving the requested streaming file from disk if determined that the requested streaming file is not in the server's memory.

20. The system of claim 17 wherein said admission controller is further operable to perform a resource availability check for the server to determine whether the server has sufficient available resources to service the new request.

21. The system of claim 20 wherein said sufficient available resources to service the new request comprises sufficient resources available so as not to overload the server.

22. The system of claim 17 wherein said admission controller is further operable to perform quality of service guarantee check for the server to determine whether acceptance of the new request will violate, at any point in the future, a desired quality of service provided by the server for any previously accepted requests.

23. A method comprising:

receiving, at a time T_{cur}, a new request for a streaming file to be served by a media server;

creating a segment-based model of the media server's memory as of time T_{cur}; and

based at least in part on the segment-based model of the media server's memory, determining whether to accept the received request for service by the media server.

24. The method of claim 23 wherein said segment-based model of the media server's memory comprises (a) identification of unique segments of streaming files previously accessed by clients of the media server and (b) identification of corresponding timestamps of most recent accesses of each unique segment.

25. The method of claim 23 wherein said determining whether to accept the received request for service by the media server comprises:

determining whether the received request can be serviced by the media server without overloading the media server.

26. The method of claim 23 wherein said determining whether to accept the received request for service by the media server comprises:

determining a cost to the server for serving the requested streaming file, wherein the cost corresponds to the amount of the media server's resources to be consumed in serving the requested streaming file.

27. The method of claim 23 wherein said determining whether to accept the received request for service by the media server comprises:

performing a resource availability check for the media server to determine whether the media server has sufficient available resources to service the new request.

28. The method of claim 23 wherein said determining whether to accept the received request for service by the media server comprises:

performing quality of service guarantee check for the media server to determine whether acceptance of the new request will violate, at any point in the future, a desired quality of service provided by the media server for any previously accepted requests.

29. Computer-executable software stored to a computer-readable medium, the computer-executable software comprising:

code for creating a segment-based model of a media server's memory; and

code for determining whether to serve a requested streaming file from the media server based at least in part on the segment-based model of the media server's memory.

30. The computer-executable software code of claim 29 further comprising:

code for receiving a request for said streaming file.

31. The computer-executable software code of claim 30 further comprising:

code, responsive to receiving said request, for determining whether to accept the request for service by the media server.

32. The computer-executable software code of claim 31 wherein said code for determining whether to accept the request for service by the media server comprises:

code for determining whether the request can be serviced by the media server without overloading the media server.

33. The computer-executable software code of claim 29 wherein said segment-based model of the media server's memory comprises (a) identification of unique segments of streaming files previously accessed by clients of the media server and (b) identification of corresponding timestamps of most recent accesses of each unique segment.

34. The computer-executable software code of claim 29 wherein said code for determining whether to serve a requested streaming file from the media server comprises:

code for determining a cost to the media server for serving the requested streaming file, wherein the cost corresponds to the amount of the media server's resources to be consumed in serving the requested streaming file.

35. The computer-executable software of claim 29 wherein said code for determining whether to serve a requested streaming file from the media server comprises:

code for performing a resource availability check for the media server to determine whether the media server has sufficient available resources to service the new request.

36. The computer-executable software code of claim 29 wherein said code for determining whether to serve a requested streaming file from the media server comprises:

code for performing quality of service guarantee check for the media server to determine whether acceptance of the new request will violate, at any point in the future, a desired quality of service provided by the media server for any previously accepted requests.

37. A cost-aware admission control system comprising:

means for receiving, at a time T_{cur}, a new request for a streaming file to be served by a media server;

means for creating a segment-based model of the media server's memory as of time T_{cur}; and

means for determining, based at least in part on the segment-based model of the media server's memory, whether to accept the received request for service by the media server.

38. The cost-aware admission control system of claim 37 wherein said segment-based model of the media server's memory comprises (a) identification of unique segments of streaming files previously accessed by clients of the media server and (b) identification of corresponding timestamps of most recent accesses of each unique segment.

39. The cost-aware admission control system of claim 37 wherein said means for determining whether to accept the received request for service by the media server comprises:

means for determining whether the received request can be serviced by the media server without overloading the media server.

40. The cost-aware admission control system of claim 37 wherein said means for determining whether to accept the received request for service by the media server comprises:

means for determining a cost to the server for serving the requested streaming file, wherein the cost corresponds to the amount of the media server's resources to be consumed in serving the requested streaming file.